Use of Ontology and Cluster Ensembles for Geospatial Clustering Analysis

Wei Gu[1], Zhilin Zhang[2], Baijie Wang[3], and Xin Wang[2]

[1] petroWeb, 1200 333-7th Avenue, Calgary, AB, Canada
victorgucanada@hotmail.com
[2] Department of Geomatics Engineering, University of Calgary
2500 University Drive NW, Calgary, AB, Canada, T2N 1N4
{zhi-lin.zhang,xcwang}@ucalgary.ca
[3] Husky Energy Inc. 707-8th Avenue, Calgary, AB, Canada
wangbaijie@gmail.com

Abstract. Geospatial clustering is an important topic in spatial analysis and knowledge discovery research. However, most existing clustering methods clusters geospatial data at data level without considering domain knowledge and users' goals during the clustering process. In this paper, we propose an ontology-based geospatial cluster ensemble approach to produce good clustering results with the consideration of domain knowledge and users' goals. The approach includes two components: an ontology-based expert system and a cluster ensemble method. The ontology-based expert system is to represent geospatial and clustering domain knowledge and to identify the appropriate clustering components (e.g., geospatial datasets, attributes of the datasets, and clustering methods) based on a specific application requirement. The cluster ensemble is to combine a diverse set of clustering results produced by recommended clustering components into an optimal clustering result. A real case study has been conducted to demonstrate the efficiency and practicality of the approach.

Keywords: Spatial analysis, Ontology, Cluster ensemble, Facility location analysis.

1 Introduction

Geospatial clustering is an important topic in spatial analysis and knowledge discovery research. It aims to partition similar objects into the same group (called a cluster) based on their similarity or connectivity in geographical space while placing dissimilar objects in different groups [1,2]. It can be used to find natural clusters (e.g., extracting the type of land use from the satellite imagery), identify hot spots (e.g., epidemics, crime, traffic accidents), and partition an area based on utility (e.g., market area assignment by minimizing the distance to customers).

Domain knowledge and users' goals play important roles during geospatial clustering [3,4,5]. The background knowledge concerning the domain described

M. Sokolova and P. van Beek (Eds.): Canadian AI 2014, LNAI 8436, pp. 119–130, 2014.

by the geospatial data is called domain knowledge. In geospatial clustering analysis, a user seeks to discover knowledge from geospatial data based on a particular goal by applying clustering methods. However, most existing clustering processes and clustering methods focus solely on the data itself without considering domain knowledge. Thus, clustering occurs at the data level instead of the knowledge level, which prevents the user from precisely understanding the clustering results and achieving his or her goals.

A few options for handling the problem seem apparent. One option is to develop new geospatial clustering methods exactly tailored for users' applications. The customized methods should consider which attributes of the geospatial data are needed and what kinds of the domain knowledge have to be exploited. Some customized clustering methods called constrained-based clustering methods, have been proposed [8,9,11]. Since they only consider limited knowledge concerning the domain and the user's goals, they are typically difficult to be reused. In particular, they usually have very restricted means of incorporating domain-related information from non-geospatial attributes.

The second option can be defined by considering the overall nature of the clustering process and building a knowledge-based system to support the integration of knowledge in the geospatial clustering process. The clustering process consists of all steps required to accomplish a clustering task given by a user. It starts from data preprocessing (including data cleaning, data integration, data selection, and data transformation), then applies clustering methods on the datasets, and finally presents the clustering results to the user [12]. Applying a clustering method is only one step of the overall process. Thus, the second option is to incorporate domain knowledge and users' goals into the clustering process, which allows an informed choice to be made from choosing the available datasets, suitable attributes of the datasets and clustering methods. However, this option can only get the best clustering results by using the existing most suitable clustering method and thus will be not helpful when none of the existing clustering method can provide good clustering results for a specific application.

The third option is to apply cluster ensembles [13,29] to geospatial clustering analysis. By applying available clustering methods to different attributes of datasets, cluster ensembles can obtain a large set of clustering results and finally combine them into a single consolidated clustering result. The result contains all information in the ensemble. However, it is time consuming to get a diverse set of clustering results. Previous research also shows that it is not always the best to include all available clustering results in the ensemble [14,28]. Thus, there is an emerging interest on reducing the number of clustering results in the ensemble.

In this paper, we propose a novel approach to geospatial clustering analysis, which combines an ontology-based expert system with a cluster ensemble method. Specifically, we first build an ontology to represent geospatial and clustering domain knowledge and then use an expert system to help identify appropriate geospatial datasets, attributes of the datasets and clustering methods for a specific application. Next, all the datasets, the attributes of the datasets and the clustering methods recommended by the ontology-based expert system

are used to produce a diverse set of clustering results. Finally, with the help of the domain knowledge in the expert system, a subset of clustering results are selected and combined into a single clustering result. The approach can perform better than existing clustering methods because of the following reasons:

First, instead of developing a new clustering method for every specific application, the approach considers knowledge reuse. The domain knowledge learned from previous applications is formalized into the ontology-based expert system and would be reused for the similar applications in the future.

Second, with the help of the domain knowledge, the approach can identify appropriate components, including appropriate datasets, attributes of datasets, and clustering methods, according to users' goals.

Third, the approach combines the clustering results produced by the appropriate clustering components and results in one best clustering result using the cluster ensemble method. It provides more comprehensive clustering results when none of the clustering results produced by the appropriate clustering components is good enough.

The rest of the paper is organized as follows. The ontology-based geospatial clustering ensemble approach is proposed in the Section 2. A case study of the approach to do facility location analysis in Alberta, Canada is presented in Section 3. Section 4 summarizes the paper and discusses the future work.

2 An Ontology-Based Geospatial Cluster Ensemble Approach

In this section, we present an ontology-based cluster ensemble approach for geospatial clustering analysis. The approach includes two components: the ontology-based geospatial clustering system and a clustering ensemble method. In the following, we start with the general workflow of the approach and then we will introduce each component in detail.

2.1 Work Flow of the Approach

As shown in Fig. 1, users' clustering goals are first sent to the ontology-based geospatial clustering system, GEO_CLUST. The GEO_CLUST identifies appropriate geospatial datasets, attributes of the datasets and clustering methods according to the goals and the domain knowledge. Then, a diverse set of clustering results are produced by applying different clustering methods to different datasets and attributes of datasets multiple times.

For a better understanding of the approach, we treat each clustering result of a geospatial clustering application as a solution to that application and plotted them into a solution space. Fig. 2 shows the changes in a solution space when applying the ontology-based geospatial cluster ensemble approach under different statuses. At status 1, only the solutions within the circle are left for the following analysis. Second, a set of clustering results are sent the cluster ensemble method. In the first step of the method, a subset of the clustering results are selected

with the criteria of high quality and diversity. The domain knowledge may be extracted from the GEO_CLUST to measure the quality. Thus at status 2, the solution space is further reduced, as shown the smaller rectangle in Fig. 2. Finally, the second step of the cluster ensemble method combines the selected clustering results into one optimal combined clustering result (as shown the black point in Fig. 2). According to the Equation (6), the combination may need the domain knowledge which could be extracted from the GEO_CLUST.

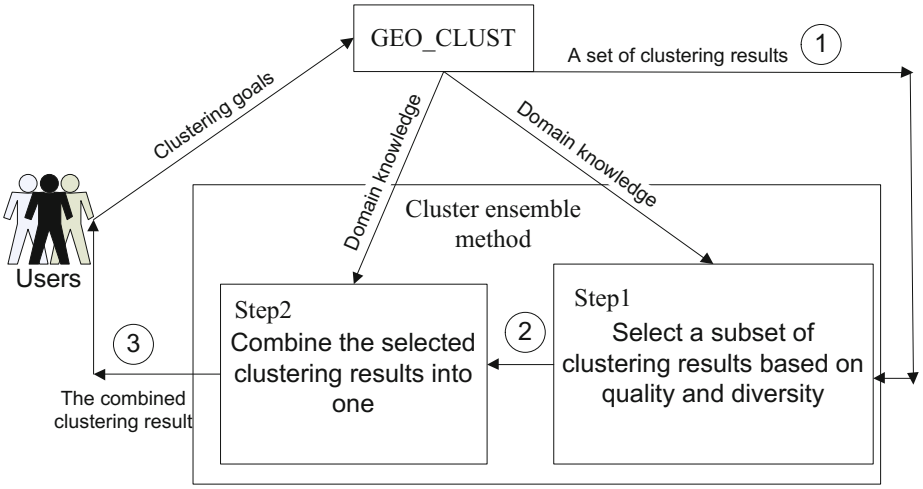

Fig. 1. Work flow of the approach

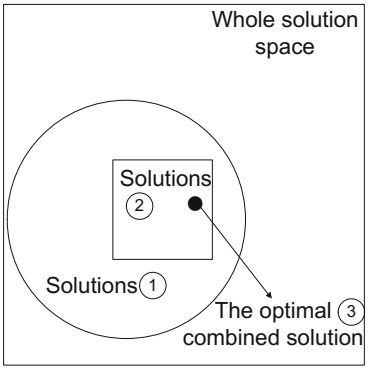

Fig. 2. Changes in solution space

2.2 An Ontology-Based Geospatial Clustering System

In this section, we present an ontology-based expert system, named GEO_CLUST, for performing geospatial clustering. The goal of the system is to

make better use of geospatial and clustering knowledge to select proper methods and datasets to achieve clustering results that better meet users' requirements. The system consists of the GeoCO ontology for geospatial clustering and the Ontology Reasoner reasoning mechanism. The GeoCO ontology is used to represent geospatial and clustering domain knowledge. The Ontology Reasoner uses classification and decomposition techniques to specify users' tasks.

An **ontology** is a formal explicit specification of a shared conceptualization [15]. It provides domain knowledge relevant to the conceptualization and axioms for reasoning with it. For geospatial clustering, an appropriate ontology must include a rich set of geospatial and clustering concepts. Therefore, it can provide a knowledge source that supplements domain experts. Since the ontology in the geospatial domain is complex and varies according to the application [18], we build GeoCO at a high generic level such that it can be extended and materialized for specific applications. The GeoCO geospatial clustering ontology has been represented in using Protg-OWL [24] and the detail information about it can be found in [4].

The structure of the GEO_CLUST system for ontology-based clustering is shown in Fig. 3. It includes five components: the Geospatial Clustering Ontology, the Ontology Reasoner, the Clustering Methods, the Data, and the Graphical User Interface (GUI). The geospatial clustering ontology component is used when identifying the clustering problem and the relevant data. Within this component, the task model specifies the data and methods that may potentially be suitable for meeting the user's goals, and GeoCO includes all classes, instances, and axioms in a geospatial clustering domain. Through classification and decomposition conducted in the Ontology Reasoner, proper clustering data and methods can be indentified from the ontology.

The system works as follows: with the system, the user first gives his or her goals for clustering through the GUI. To be able to find proper data and clustering methods in ontology, the goal needs formalizing as a task instance. A task instance describes the specific problem to be solved. An example of a user's goal is to identify the best locations for five hospitals in Alberta. A task instance "determine best locations of five hospitals in Alberta" is created. The task instance is refined in the task model by using Ontology Reasoner and the refined elementary sub-tasks are used to search the domain ontology. For the example above, one of elementary tasks of "determine the best locations of five hospitals in Alberta" is "to find population data in Alberta," which can be implemented through database queries. The results of these queries identify the proper clustering methods and the appropriate data sets. Based on these results, clustering is conducted. The clustering results can be used for statistical analysis or be interpreted using the task ontology and the domain ontology.

In the system, the Ontology Reasoner is used to reason about knowledge represented in the ontology. In this component, classification are applied to detect the most specialized task node in a tree structure that the task instance belongs to, where the tree is used to organize all tasks hierarchically. Specifically, each task instance is classified according to its data and constraints, and thus the

sub-task best fitting the characteristics of the task instance can be selected for the following processes. Decomposition describes the process of decomposing a task into simpler but more elementary subtasks, which presents a problem-solving strategy for tackling the task with a list of sub-tasks and operators (sequence, choice, iteration).

The input of the reasoner is the user's goals, and the output is a set of appropriate geospatial clustering methods and datasets. The reasoner performs the following steps. First, it builds a task instance [25] to associate the reasoning with the geospatial clustering ontology and the user's requirements. For the above example, the task instance "determine best locations of five hospitals in Alberta" is created based on a Partitioning-Clustering task in the task model, because the final results of clustering are to form five population clusters assigned to individual hospitals. Second, each available geospatial clustering method described in the ontology is considered either as an elementary task (which is accomplished by a simple primitive function) or as a complex task (which is accomplished via a task decomposition method represented in some problem solving strategy). The task "determine best locations of five hospitals in Alberta" is a complex task because we cannot solve the task by simply calling an existing primitive function. For the task, we first need to find the proper data, such as Alberta population data, and then identify the proper partitioning clustering method based on the characteristics of the data and the user's specification of "best locations". So the task needs to be decomposed. Finally each complex task is recursively decomposed into elementary sub-tasks [26]. The detailed description about the Ontological Reasoner component in GEO_CLUST is in [4], including task model, inference engine, and classification algorithm.

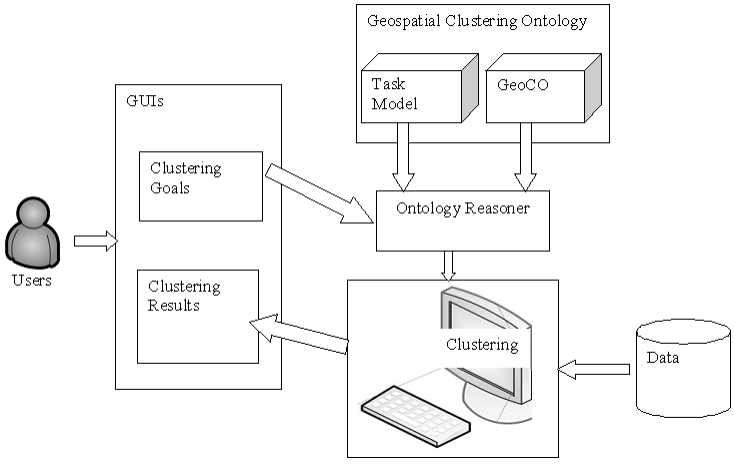

Fig. 3. The GEO_CLUST system for ontology-based clustering

2.3 A Cluster Ensemble Method

Given a set of clustering results, the cluster ensemble method used in the approach aims to select a subset of the clustering results and then combines them into a new clustering result which is better than the best result in the given result set. We extend cluster ensemble selection method in [14] by cooperating with the domain knowledge. The method includes two steps.

Step one: select a subset of clustering results with high quality and diversity, where quality measures the accuracy of clustering results in the subset and diversity measures the difference among the clustering results. According to the research did by Fern and Lin [14], selecting only a subset of clustering results based quality and diversity could improve the accuracy of the final ensemble result as well as reducing the execution time. Particularly, in this paper, given a set of clustering results C (i.e., $C = \{C_1, C_2, ..., C_r\}$) and a subset $C' \in C$, the way to measure the quality of C' is separated into two conditions:

(1) Without the domain knowledge. The quality of $(C_i \in C')$ is defined as the similarity between it and the other clustering results in C (as shown in Equation (1)) and the quality of C' is defined as the sum of the quality of all the clustering results in it (as shown in Equation (2)).

$$Quality(C_i) = \sum_{k=1}^{r} NMI(C_i, C_k) \tag{1}$$

$$Quality(C') = \sum_{C_i \in C'} Quality(C_i) \tag{2}$$

Where $NMI(C_i, C_k)$ is the normalized mutual information[1] between clustering C_i and C_k. We adopt the Equation 3 in [13] to estimate the NMI value between two clustering results. According to [13], if two clusterings are completely independent partitions, their NMI value is 0, vice versa. Thus, the larger is the value, the higher is the quality.

(2) With the domain knowledge. The quality of C_i $(C_i \in C')$ is measured by the external objective function according to the domain knowledge, and the quality of C' is defined as the sum of the quality of all the clustering results in it (as shown in Equation (3)).

$$Quality(C') = \sum_{C_i \in C'} EF(C_i) \tag{3}$$

Where $EF(C_i)$ is the external objective function value of C_i. For instance, when applying geospatial clustering analysis for the facility location planning [7], all the demand nodes in the target region are clustered into different groups and the demand nodes in each group are served by one facility. According to the domain knowledge in facility location planning, the external objective function

[1] Mutual information is a symmetric measure to quality the statistical information shared between two distributions.

value of a clustering result is the total travelling distance from the demand nodes to their assigned facilities.

The diversity of C' is defined as the sum of all pairwise similarities in the set (as shown the Equation (4)). The lower the value, the higher is the diversity. We measure the diversity as follows because it has been proved to efficiently affect the cluster ensemble performance [14].

$$Diversity(C') = \sum_{i \neq j, C_i, C_j \in C'} NMI(C_i, C_j) \tag{4}$$

Since high quality and diversity are two objectives to be achieved during the clustering result selection, we treat it as a bi-objective optimization problem [6] and solve it by using a bi-objective function. Specifically, given a set of clustering results, a subset of clustering results C' is selected from the whole set that minimizes the value of $BOF(C')$ in the following.

$$BOF(C') = \alpha Quality(C') + (1 - \alpha) Diversity(C') \tag{5}$$

In the Equation (5), α is defined as a co-efficient for balancing the quality objective and diversity objective, which is within [0,1]. The parameter α is a constant value to control the weight of each objective.

Selecting a subset of solutions to minimize the value of $BOF(C')$ is a NP-hard problem. In the research, we adopt a greedy procedure to perform the selection. It begins with a C' that only contains the single solution of the lowest linear normalized cost value and then incrementally adds one solution at a time into C' to minimize the value of $BOF(C')$. The procedure stops when the size of C' reaches the predefined number.

Step two: combine the selected clustering results into one optimal combined clustering result. The optimal clustering result C_{opt} should reach the objective function (Equation (6)), which asks for the optimal result has maximal mutual information with other selected clustering results and achieves higher external objective function value if domain knowledge is applied. We adopt the greedy optimization approach in [13] to solve the Equation (6) by defining our objective function $\Gamma(C_{opt})$.

$$\Gamma(C_{opt}) = \arg\max_{C_{opt}} \sum_{C_i \in C'} \left(\beta EF(C_{opt}) + (1 - \beta) NMI(C_{opt}, C_i)\right) \tag{6}$$

Where C_{opt} is a single labeling of clustering results which maximizes $\Gamma(C_{opt})$.

3 Case Study

In this section, we apply the approach to a real case study, finding the best locations for breast cancer screening clinics in Alberta (AB), Canada.

A population-based program to increase the number of Alberta women screened regularly for breast cancer was implemented in 1990 and today the Alberta Breast Cancer Screening Program (ABCSP) recommends Alberta women

between the ages of 50 and 69 have a screening mammogram at least once every two years [10]. A key challenge is to determine the best locations for screening clinics to minimize the average demand-weighted distance from demand nodes to their assigned clinics. In this case, we perform the approach to separate the whole province into 26^2 screening clusters and for each cluster find a suitable location for building a clinic to serve the people within it.

In the approach, the first step is to find appropriate datasets, attributes of datasets, and clustering methods by sending the clustering goal to the ontology-based geospatial clustering system. The user interface and parameter setting of the ontology-based system can be found in [4]. Here we only list the results as below:

- Dataset: Census dataset in AB
- Attributes of datasets: women from 50 to 69 in DA level
- Similarity Measurement: Euclidean distance
- Clustering method: Capability K-MEANS

The dataset is the 2006 Canadian census data in AB. Estimates of the screening population (Alberta women aged 50 to 69 years) were derived from census data at the Dissemination Area (DA) level [16]. There are 327830 women within the target age in Alberta. A total of 5180 DAs were used in the research. Their values range from 0 to 920. The system adopts Euclidean distance to measure the similarity among the locations. The recommended clustering method is Capability K-MEANS [17]. In order to meet the overall demands in the province, the capacity of each clinic is set to 15,000. We generate 30 clustering results by applying Capability K-MEANS with different initializations.

The second step is to select a subset of clustering results based on quality and diversity. The Equation (3) in 3.1 is chosen to measure the quality. The external objective function value is the total travelling distance from the DAs to their assigned clinics. The diversity is measured by the Equation (4) in 3.1. α in the Equation (5) is set to 0.5. The number of the selected solutions is set to 5.

In the third step, the selected 5 results are combined into one optimal result. Since the case has a clear clustering goal, minimizing the average demand-weighted distance from demand nodes to their assigned clinics, the way to choose the optimal result is only based on the goal and thus the value of β in the Equation (6) is set to 1. Fig. 4 shows the clinic locations in the optimal result. We compare the optimal result with the one produced by the Interchange algorithm [27], the most popular algorithm used in facility location planning. The average demand-weighted distance are 24.12 km for the optimal result and 26.64 for the one produced by the Interchange algorithm. In this case, the approach achieves better results than by simply picking a standard facility location solution approach.

Finally, the knowledge obtained from this case study can be incorporated into some other applications in the future. For example, the same results can be used for planning the locations of pharmacies in the communities.

[2] The number of current cancer care facilities in Alberta is 26.

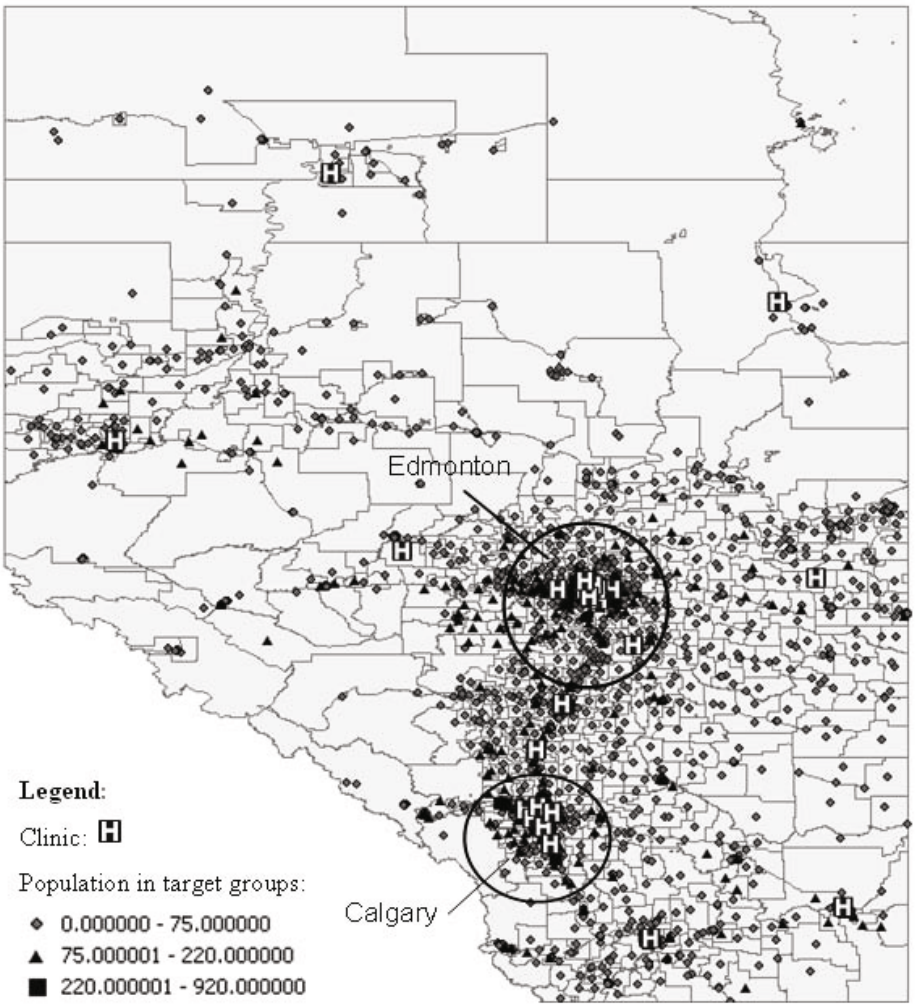

Fig. 4. The The optimal clinics distribution in AB

4 Conclusion and Future Work

In this paper, we present an ontology-based cluster ensemble approach to produce good clustering results for geospatial applications. The approach includes two components: an ontology-based expert system for formalizing the domain knowledge in the geospatial clustering, and a cluster ensemble method for combing good clustering results into an optimal result. To our best knowledge, it is the first research work to combine geospatial ontology and cluster ensembles for geospatial clustering analysis. The real case study did in Alberta, Canada shows that it is practical to combine the ontology and the cluster ensembles together

for geospatial analysis. In the future, we will investigate how to further improve clustering results by integrating clustering ensemble with domain knowledge in theory and apply the presented approach on more application scenarios.

References

1. Ng, R., Han, J.: Efficient and Effective Clustering Method for Spatial Data Mining. In: Proc. of 20th International Conference on Very Large Data Bases, pp. 144–155. Morgan Kaufmann, San Francisco (1994)
2. Shekhar, S., Chawla, S.: Spatial Databases: A Tour. Prentice Hall (2003)
3. Graco, W., Semenova, T., Dubossarsky, E.: Toward knowledge-driven Data Mining. In: Proc. of International Workshop on Domain Driven Data Mining at 13th ACM SIGKDD, pp. 49–54. ACM, New York (2007)
4. Wang, X., Gu, W., Ziebelin, D., Hamilton, H.: An Ontology-based Framework for Geospatial Clustering. International Journal of Geographical Information Science 24, 1601–1630 (2010)
5. Wang, X., Hamilton, H.J.: Towards an Ontology-based Spatial Clustering Framework. In: Kégl, B., Lee, H.-H. (eds.) Canadian AI 2005. LNCS (LNAI), vol. 3501, pp. 205–216. Springer, Heidelberg (2005)
6. Mitropoulos, P., Mitropoulos, I., Giannikos, I., Sissouras, A.: A Biobjective Model for the Locational Planning of Hospitals and Health Centers. Health Care Management Sci. 9, 171–179 (2006)
7. Liao, K., Guo, D.: A Clustering-Based Approach to the Capacitated Facility Location Problem. Trans GIS 12, 323–339 (2008)
8. Prabakara Raj, S.R., Ravindran, B.: Incremental Constrained Clustering: A Decision Theoretic Approach. In: Li, J., Cao, L., Wang, C., Tan, K.C., Liu, B., Pei, J., Tseng, V.S. (eds.) PAKDD 2013 Workshops. LNCS, vol. 7867, pp. 475–486. Springer, Heidelberg (2013)
9. Wang, X., Rostoker, C., Hamilton, H.J.: Density-based Spatial Clustering in the Presence of Obstacles and Facilitators. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 446–458. Springer, Heidelberg (2004)
10. Alberta Breast Cancer Screening Program website, `http://www.cancerboard.ab.ca/abcsp/program.html`
11. Thiago, F.C., Eduardo, R.H., Joydeep, G.: A Study of K-Means-based Algorithms for Constrained Clustering. J. Intelligent Data Analysis 17, 485–505 (2013)
12. Han, J.W., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2011)
13. Strehl, A., Ghosh, J.: Cluster Ensembles A Knowledge Reuse Framework for Combining Multiple Partitions. Machine Learning Research 3, 583–617 (2002)
14. Fern, X.Z., Lin, W.: Cluster Ensemble Selection. Journal of Statistical Analysis and Data Mining 1, 128–141 (2008)
15. Gruber, T.R.: A Translation Approach to Portable Ontologies. Knowledge Acquisition 5, 199–220 (1993)
16. Data quality index for census geographies, `http://www12.statcan.ca.ezproxy.lib.ucalgary.ca/census-recensement/2006/ref/notes/DQ-QD_geo-eng.cfm`
17. Ng, M.K.: A Note on Constrained k-means Algorithms. Pattern Recognition 33, 515–519 (2000)

18. Fonseca, F., Egenhofer, M., Agouris, P., Cmara, G.: Using Ontologies for Integrated Geographic Information Systems. Transactions in GIS 6, 231–257 (2002)
19. Maedche, A., Zacharias, V.: Clustering Ontology-based Metadata in the Semantic Web. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 348–360. Springer, Heidelberg (2002)
20. Worboys, M.F.: Metrics and Topologies for Geographic Space. In: Advances in Geographic Information Systems Research II: International Symposium on Spatial Data Handling (1996)
21. Egenhofer, M.J., Clementini, E., di Felice, P.: Topological Relations between Regions with Holes. International Journal of Geographical Information Systems 8, 129–142 (1994)
22. Papadias, D., Egenhofer, M.: Hierarchical Spatial Reasoning about Direction Relations. GeoInformatica 1, 251–273 (1997)
23. Egenhofer, M.J., Franzosa, R.D.: Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems 5, 161–174 (1991)
24. Protg web site, `http://protege.stanford.edu/index.html`
25. Crubzy, M., Musen, M.: Ontologies in Support of Problem Solving. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 321–341. Springer, Heidelberg (2004)
26. Parmentier, T., Ziébelin, D.: Distributed Problem Solving Environment Dedicated to DNA Sequence Annotation. In: Fensel, D., Studer, R. (eds.) EKAW 1999. LNCS (LNAI), vol. 1621, pp. 243–258. Springer, Heidelberg (1999)
27. Teitz, M.B., Bart, P.: Heuristic methods for estimating the generalized vertex median of a weighted graph. Oper. Res. 16, 955–961 (1968)
28. Naldi, M.C., Carvalho, A.C.P.L.F., Campello, R.J.G.B.: Cluster Ensemble Selection Based on Relative Validity Indexes. Data Mining and Knowledge Discovery 27, 259–285 (2013)
29. Sarumathi, S., Shanthi, N., Santhiya, G.: A Survey of Cluster Ensemble. International Journal of Computer Applications 65, 8–11 (2013)